The Chinese Room Reversed: Meaning, Subjectivity and Artificial Intelligence

Edward Grabczewski
Neath, United Kingdom
20 August 2026
egrabczewski@gmail.com

Abstract: This paper revisits the debate between Alan Turing's behavioural conception of machine intelligence and John Searle's Chinese Room argument. Drawing on the author's experience as a professional software engineer, it contrasts traditional programming, in which the programmer explicitly creates the rules followed by the machine, with contemporary interactions with large language models, in which that relationship can appear partially reversed. A recent experience using ChatGPT to diagnose and modify legacy C software is used to question where understanding should be located within an intelligent system. The paper argues that semantics may depend not merely upon symbol manipulation but upon the construction and manipulation of a subjective model grounded in interaction with an external world. It further considers the Systems Reply to Searle and suggests that perception, action, memory, feedback and causal engagement with the environment may be central to the emergence of meaning. The conclusion is that intelligence should not necessarily be judged according to whether another system's internal meanings are accessible in human terms.

Keywords: Artificial intelligence; Chinese Room; Alan Turing; John Searle; machine understanding; semantics; symbol grounding.

1. Introduction

Alan Turing's paper *Computing Machinery and Intelligence* (Turing, 1950) begins with the deceptively simple question: "Can machines think?" Rather than attempting to define either "machine" or "thinking" directly, Turing proposed replacing the question with a practical test: the imitation game. If, after a period of text-based questioning, a human interrogator could not reliably distinguish a machine from a human being, Turing suggested that it would be reasonable to regard the machine as exhibiting intelligence.

Turing even ventured a quantitative prediction. He believed that within about fifty years computers would become sufficiently capable that, after five minutes of questioning, an average interrogator would have no more than a 70 per cent chance of correctly identifying which participant was the machine. In other words, the machine would fool the interrogator about 30 per cent of the time.

Modern large language models have made that prediction look remarkably prescient. Systems such as ChatGPT can sustain conversations that are fluent, coherent, knowledgeable and responsive across an extraordinary range of subjects. They undoubtedly make mistakes, but humans make mistakes too. We misremember facts, misunderstand questions, contradict ourselves, make computational errors and sometimes express unjustified confidence. The fact that an artificial intelligence makes mistakes therefore cannot, by itself, distinguish it from a

human intelligence. Indeed, excessive accuracy, consistency and fluency might sometimes make a machine appear less human rather than more so.

2. Searle and the Chinese Room

Turing's behavioural approach was famously challenged by the philosopher John Searle in his Chinese Room argument (Searle, 1980). Searle asks us to imagine a person who does not understand Chinese sitting in a room. Chinese symbols are passed into the room, and the person is supplied with a book containing detailed instructions, written in a language he does understand, telling him how to manipulate those symbols and which symbols to return. By following the rules sufficiently well, the person may produce answers indistinguishable from those of a native Chinese speaker.

Yet the person inside the room still does not understand Chinese.

For Searle, this demonstrates a fundamental distinction between syntax and semantics. A formal system may manipulate symbols entirely according to rules while possessing no understanding of what those symbols mean. Correct symbol manipulation, however sophisticated, does not necessarily produce meaning. His argument therefore attacks the strong claim that the execution of the appropriate computer program is, in itself, sufficient for understanding.

I have always found this distinction important. Symbol manipulation is not automatically semantics.

At the same time, when I first encountered Searle's Chinese Room argument, one feature of the thought experiment struck me as peculiar. The person in the room may be manipulating symbols mindlessly, but the book of rules from which he is working did not write itself. It was created by an intelligent being.

3. The Programmer as Author of the Rule Book

For most of the history of computing, that observation seemed particularly significant. Computer programs were written by human beings. A programmer understood a problem, devised a procedure for solving it, expressed that procedure as an algorithm and encoded it in a programming language. The computer then followed those instructions mechanically.

I was a professional software engineer for twenty-eight years, and for eight of those years I worked professionally as a C programmer. I therefore spent a substantial part of my working life doing precisely what the traditional model of computing assumes: understanding a problem and translating that understanding into formal instructions which a computer could execute without possessing any understanding of its own.

One of the most intellectually demanding programming tasks of my career involved recreating part of the VAX/VMS command-line system in C for UNIX. The development process required a form of bootstrapping: I first had to construct a simple version of a translation tool, including some of its initial data by hand, so that this primitive tool could then be used to help construct the more sophisticated version that would eventually replace it.

The process was mind-bending because I had to keep several levels of abstraction in my head at once. Yet however complicated the machinery became, there was never any doubt about where the higher-level understanding resided. I had conceived the strategy. I understood why the first primitive tool was needed, why part of it had to be constructed manually, and how it would subsequently be used to create the full system. The computer simply carried out the formal operations I had designed.

This experience is one reason why Searle's Chinese Room originally seemed so persuasive to me. I knew from years of professional programming what it was like to be the intelligence that created the "rule book". The machine followed the rules; I understood their purpose.

4. The Chinese Room Reversed

That is precisely why my much later experience with ChatGPT-5.5 while porting REQUIEM was so striking. For the first time, I sometimes found the relationship reversed. ChatGPT was devising the diagnostic strategy, telling me what commands to execute and interpreting the results that I returned. I knew what it was like to write Searle's rule book. Now, unexpectedly, I sometimes found myself in the position of the person following it.

REQUIEM is a relational database management system developed by Mike P. Papazoglou and colleagues (Papazoglou, 1989) and originally written in C for the 32-bit UNIX systems of the time. I was attempting to compile and run it on a modern Apple iMac. Unsurprisingly, the process produced numerous compiler warnings, compatibility problems and runtime faults. I then decided to explore ChatGPT's ability to help diagnose and resolve them.

ChatGPT would ask me to inspect particular files or regions of source code. It would tell me to run commands using tools such as *grep* and *sed*, and I would paste the resulting text back into the conversation. It would analyse what I had returned, infer what might be wrong, suggest changes to the source and ask me to compile the program again. When the fault required runtime investigation, ChatGPT introduced me to `lldb`, the debugger available on the system.

Although I had many years of professional programming experience, including substantial experience of C, I had no knowledge of `lldb`. ChatGPT did.

It told me which commands to enter, which parts of the resulting output were significant, and what should be investigated next. I executed the commands and returned the results. ChatGPT interpreted them and modified its diagnostic strategy accordingly.

At some point I realised, with considerable amusement, that Searle's arrangement had been partially reversed.

In the Chinese Room, Searle places a human being in the role traditionally occupied by a computer: the human mechanically executes rules without understanding the symbols being manipulated. During my debugging session, I sometimes occupied something remarkably similar to that role. ChatGPT instructed me to type commands into a debugger whose operation I did not understand, retrieve particular pieces of information and return the resulting symbols for further interpretation.

It was ChatGPT, rather than me, that appeared to possess the relevant working model of `lldb`.

Of course, I was not literally mindless. I understood programming, C, compilers and the general problem we were solving. I could question ChatGPT's suggestions or refuse to carry them out. But in relation to the debugger itself, the distribution of knowledge was nevertheless striking. ChatGPT knew which commands were relevant, what they would reveal, how their output should be interpreted and what should be done next. I did not.

This raises an interesting question about what we mean when we say that something "understands" a tool.

If an entity can select an appropriate debugger command in response to a particular fault, anticipate what information that command should reveal, correctly interpret the resulting output, revise its hypothesis accordingly and choose a useful next action, in what sense does it fail to understand the debugger?

One could always insist that it merely manipulates symbols. But if every successful application of knowledge can be redescribed at a sufficiently low level as symbol manipulation, then the same description can ultimately be applied to human brains.

5. Meaning, Subjectivity and Symbol Grounding

Searle himself was well aware that the brain is a physical system composed of neurons. Individual neurons do not understand Chinese, C programming or philosophy. Somehow, however, the activity of vast numbers of neurons, organised into an appropriate biological system and connected continuously to sensory inputs and bodily outputs, gives rise to meaning, understanding and intelligence.

This, in my view, is where the central issue lies.

The brain is not simply a network of neurons operating in isolation. It is a network embedded within a living organism that interacts continuously with an external world. It receives sensory information, acts upon its environment, experiences the consequences of those actions, remembers previous interactions and continually constructs and modifies an internal model of reality.

I believe that this interaction with the world, and the creation and manipulation of a subjective model of that world, is fundamental to intelligence.

Meaning does not arise simply because one symbol is associated with another symbol. It arises because internal states acquire significance for the system that possesses them.

This is closely related to what philosophers and cognitive scientists call the symbol-grounding problem (Harnad, 1990). If every symbol gets its meaning only through relationships with other symbols, meaning appears endlessly deferred. A dictionary cannot ultimately define every word only by reference to other words unless some of those words are grounded in experience. At some point there must be a connection between representation and the world represented.

Human beings acquire such grounding naturally. A child does not learn the meaning of "hot" merely by consulting a dictionary. The word becomes connected to perceptions, sensations,

actions, warnings, memories and expectations. “Hot” becomes part of a model of the world that has significance to the child.

But I would go further than simply requiring sensory grounding. I think the model is necessarily subjective.

By “subjective”, I do not mean merely emotional or arbitrary. I mean that the world is represented from the standpoint of the particular intelligence that encounters it. The meaning of a stimulus depends upon the organism or system receiving it.

Consider something as simple as an amoeba. An amoeba responds to chemical gradients, temperature, contact and other environmental conditions. We should be cautious about attributing anything resembling human consciousness to such an organism, but nevertheless these stimuli have significance for it. Some conditions lead it towards nutrients; others cause avoidance. The organism possesses mechanisms through which differences in its environment are interpreted in ways relevant to its continued existence.

Those signals need not mean anything to a human observer in the way that they mean something to the amoeba.

The same principle becomes more obvious when we consider highly intelligent animals.

Dolphins communicate with one another through complex vocalisations and behavioural signals. Humans have studied dolphin communication extensively, yet we still cannot simply enter their communicative world and translate it as though it were another human language. Nevertheless, it would be extraordinary to conclude that dolphin signals therefore possess no meaning.

They clearly possess significance to dolphins.

The fact that another species cannot penetrate that significance does not make the significance unreal.

This leads me to a principle which I think is important in considering artificial intelligence:

Meaning need not be human-readable in order to be meaning.

If meaning is subjective to the intelligence that possesses it, then an outside observer cannot reasonably demand that another kind of intelligence represent the world in the same way that a human being does before acknowledging that it possesses understanding.

We do not require dolphins to explain their concepts in English before allowing that they understand one another. We do not require an amoeba to provide a philosophical account of its response to a chemical stimulus. We infer meaningful internal distinctions from the organised way in which an organism responds to its environment.

This is why I find the question of ChatGPT’s apparent understanding particularly interesting.

When ChatGPT used `lldb` during the REQUIEM debugging process, its apparent model of the tool was not my model. I had none. Yet ChatGPT could manipulate information about the debugger in a way that was coherent, purposeful and productive. It related commands to anticipated consequences, interpreted returned information and altered its subsequent behaviour accordingly.

If understanding consists in possessing and manipulating a model in which distinctions have significance to the system using them, then there is at least a serious question as to whether this should count as a form of understanding.

The important question would no longer be whether ChatGPT understands `lldb` in the same subjective manner that I would understand it after learning the debugger myself. The question would be whether ChatGPT possesses an internal organisation that plays for it the functional role that an internal model plays for me.

6. The Systems Reply and the External World

This also changes how I view the Chinese Room.

Searle asks us to focus on the individual inside the room. That person does not understand Chinese. Therefore, he argues, executing the program does not produce understanding.

But perhaps the individual is simply the wrong level at which to look.

A neuron in my brain does not understand the sentence I am writing. Nor does an individual neuron know what a debugger is. Yet it would be absurd to conclude from this that I cannot understand these things.

Likewise, the person in the Chinese Room may simply be one component of a larger system.

This is essentially the well-known "Systems Reply" to Searle: although the person manipulating the symbols does not understand Chinese, perhaps the whole system — person, rule book, memory, symbol store and processes taken together — does.

Searle rejected this reply. His response was that enlarging the system does not solve the problem of meaning. Even if the man in the Chinese Room memorised the entire rule book, all the symbol tables and every procedure, so that the whole system existed inside his own head, he would still be manipulating Chinese symbols without understanding Chinese. For Searle, adding more components, more rules or greater organisational complexity therefore cannot by itself transform syntax into semantics; computation alone is not sufficient for genuine understanding.

But this raises a further question.

What if the "system" includes not only the rule book and symbol manipulation, but also inputs from the external world, outputs to it, memory, feedback, learning and the ability to alter its internal states in response to experience?

The system would no longer merely be manipulating symbols in isolation. Its internal states would be causally connected with things outside itself. It would receive information from the world, act upon that world, observe the consequences of its actions and modify its internal organisation accordingly. Surely this is one of the characteristic features of living systems on our planet.

In my view, this interaction may be crucial. Semantics may arise when a system constructs and manipulates a subjective model of the world through its encounters with it. Inputs and outputs

are therefore not peripheral additions to intelligence; they may be part of what grounds representation in meaning.

This also suggests a difficulty with Searle's insistence that the man could simply internalise the entire system. A biological mind is not merely a self-contained rule book operating inside a skull. The brain is continuously connected with a body and an environment. Its organisation and the relations among its components may themselves be causally important, but so too may the continuing relationship between the whole system and the world outside it.

7. Modern AI and the Missing Rule Book

Modern artificial intelligence makes this question considerably more interesting than it might once have appeared, because the "rule book" itself has also changed.

Traditional computer programs consisted largely of explicit instructions written by programmers. If a program responded correctly, we could usually identify the source of the behaviour: somebody had designed the algorithm.

Large language models are profoundly different in this respect. They remain computational systems, and Searle could therefore insist that they are still ultimately engaged in formal operations. But unlike conventional programs, their detailed responses are not specified in advance by a programmer through an explicit set of rules. Their behaviour emerges from a trained model whose internal organisation has been shaped by learning rather than directly written line by line.

No programmer explicitly wrote a rule saying what ChatGPT should do when a retired software engineer attempts to port REQUIEM and encounters a particular compiler warning. Nobody wrote a specific instruction telling it which `lldb` command to recommend in response to a particular runtime fault. Nobody programmed the particular sequence of questions, hypotheses and diagnostic steps that emerged during our conversation.

The behaviour arose from a trained model.

Human beings designed the architecture. Human beings developed the learning algorithms. Human beings assembled or selected training material and constructed the computational infrastructure. But the particular knowledge and behaviour expressed during an interaction were not individually programmed in the traditional sense.

The Chinese Room's rule book has therefore become difficult to locate.

The system now looks more like this:

> **human designers → learning process → trained model → generated instructions → human operator → conventional computer**

During my REQUIEM work there was also a continuing feedback loop:

> **ChatGPT proposes an action → I execute it → the computer produces output → I return the output → ChatGPT interprets it → ChatGPT proposes another action**

That begins to resemble something closer to perception and action than the static Chinese Room.

The present limitation, however, is important.

ChatGPT does not ordinarily possess the continuous autonomous sensory and physical relationship with its environment that a human being or animal possesses. During our debugging session, I acted as much of its sensory and motor apparatus. It could not independently inspect my computer, type the commands, observe the debugger and alter the files. I supplied selected information from the external world and carried out the actions it proposed.

In that sense, I temporarily functioned rather like its hands and eyes.

But this limitation may not be permanent, and the philosophical significance of removing it would be considerable.

If artificial systems eventually possess continuing channels of perception and action, persistent memory, feedback, learning and self-correction, then the comparison with biological intelligence becomes more difficult. The question will no longer be simply whether a program can manipulate linguistic symbols convincingly.

It will be whether an artificial system can develop a persistent model of a world, distinguish itself from that world, interact with it, alter its internal model in response to experience and use that model to predict and act.

For me, that is much closer to what intelligence actually consists of.

8. What Counts as Understanding?

I would therefore express my position this way:

Understanding occurs when an intelligent system constructs and manipulates a subjective model in which distinctions have significance for that system. Meaning is grounded in the relationship between the system and the world it encounters. It need not be accessible to another intelligence, nor need it take the same form as human understanding.

This does not settle the question of consciousness.

There may still be an important distinction between information having functional significance within a system and there being something it actually feels like to be that system. An artificial intelligence might construct an extraordinarily sophisticated model of its environment while possessing no subjective experience in the phenomenal sense.

That is a profound problem, but it is also a problem that exists when we consider other biological organisms.

I cannot directly experience the subjective world of a dolphin. I cannot know what, if anything, it feels like to be an amoeba. Indeed, strictly speaking, I cannot directly experience the consciousness of another human being. I infer other minds from behaviour, structure, similarity and interaction.

This is why I am reluctant to define intelligence solely from the standpoint of an external human observer.

We should be particularly careful about assuming that human consciousness and human semantics provide the only possible standard by which meaning can exist.

A dolphin's world is a dolphin's world.

An amoeba's environment has significance in whatever primitive sense its biological organisation permits.

An artificial intelligence's internal world, if such a thing develops, may likewise be unlike ours.

If we insist that we will recognise intelligence only when another system represents reality exactly as we do, we risk making the definition of intelligence circular: intelligence becomes whatever resembles human intelligence sufficiently closely.

Turing avoided this problem by concentrating upon behaviour. Searle correctly reminded us that behaviour and symbol manipulation do not automatically establish semantics. But I think a third element is necessary: the relationship between an intelligent system, its internal model and the world with which it interacts.

Turing asks:

Can its behaviour be distinguished from that of an intelligent human?

Searle asks:

Does the manipulation of symbols amount to genuine understanding?

I would add:

Does the system possess and manipulate a model of a world in which things have significance from its own standpoint?

The three questions are related, but they are not identical.

Modern artificial intelligence has made this distinction increasingly important.

9. Conclusion

The original Turing Test now seems comparatively easy. Machines can already produce language sufficiently human-like to deceive human judges under appropriate conditions. That does not prove consciousness, but it greatly weakens the claim that sophisticated linguistic behaviour is uniquely human.

Searle's deeper challenge therefore remains: does the system understand what it says?

My experience with ChatGPT makes me less willing to answer that question simply by saying that it manipulates symbols. During the REQUIEM debugging process, ChatGPT recognised what particular debugger commands were for, interpreted their output, connected observations with hypotheses, planned further investigations and revised those plans according to evidence.

Calling all of this "symbol manipulation" is technically possible. But describing my own reasoning in terms of neuronal firings is technically possible too. Neither description necessarily tells us whether understanding exists at the higher level of organisation.

Perhaps this is ultimately the weakness in the Chinese Room.

The man in Searle's room does not understand Chinese. But neither does an individual neuron in the brain of a Chinese speaker.

The relevant question may not be whether the component understands.

It may be whether the organised system does.

And the organisation and relationships between components — together with the system's continuing relationship with the external world — may themselves be precisely where the capacity for meaning arises.

This brings me back to the reversal I experienced while working on REQUIEM.

For much of my professional life I wrote instructions for computers. As a C programmer, I constructed explicit sequences of operations and expected machines to carry them out mechanically. I knew what it was like to be the intelligence that conceived the strategy and wrote the rules.

Decades later, I found myself sitting at a computer while an artificial intelligence told me what commands to type.

John Searle had placed a human being inside a room and asked him to behave like a computer in order to demonstrate that computation was not understanding. I unexpectedly found myself doing something remarkably similar — except that the computer was now outside the room, telling me which symbols to manipulate.

That experience does not prove that ChatGPT is conscious. It does not finally defeat Searle's distinction between syntax and semantics. Nor does it demonstrate that artificial and biological intelligence are equivalent.

But it does suggest that the conceptual landscape has changed.

The simple picture in which intelligence belongs to the human programmer while the computer merely obeys is no longer adequate. We are beginning to build systems whose detailed behaviour was not explicitly written by any individual human being, which can construct and manipulate extraordinarily complex internal relationships and which can sometimes direct human beings through intellectual tasks that the humans themselves do not know how to perform.

The question Turing asked has therefore evolved.

We began by asking whether a machine could imitate an intelligent human.

Searle forced us to ask whether imitation implied understanding.

The next question may be more difficult still:

If an artificial system develops a model of the world that has significance within its own processes, interacts with that world, uses its model to predict and act, and possesses forms of understanding that may not be directly accessible to us, on what principled basis could we say that its meanings are not meanings simply because they are not ours?

Acknowledgements

The author acknowledges the assistance of OpenAI's ChatGPT in discussing and developing the arguments, synthesising the source material, and drafting and editing the manuscript. The author retains responsibility for the intellectual content, interpretations and conclusions of the paper.

References

Harnad, S. (1990). *The symbol grounding problem*. Physica D, 42, 335–346.

Papazoglou, M. P., & Valder, W. (1989). *Relational Database Management: A Systems Programming Approach*. Prentice Hall. ISBN 13-771866-7

Searle, J. R. (1980). *Minds, brains, and programs*. Behavioral and Brain Sciences, 3(3), 417–424.

Turing, A. M. (1950). *Computing machinery and intelligence*. Mind, 59(236), 433–460.